Review of ”Earth system data cubes unravel global multivariate dynamics” by Mahecha et al. (esd-2019-62)

November 15, 2019

In this article Mahecha et al. present the concept of data cubes to handle the growing body of Earth system data and introduce the computing interface ”Earth system data lab” (ESDL) as a cloud-based solution. In the introduction the authors describe different data sources and variables of the Earth system, the hurdles of using the data, and illustrate the data cube approach as solution to the problem. Then, they delve into the concept and definition of data cubes, provide a generic description and mathematical formulations including how to apply customized operations on data cubes. In this context, Mahecha et al. explain the detailed implementation of the data cube approach in the ESDL project and depict its representation and processing of the various data streams. The authors showcase three example studies to demonstrate the functioning and usefulness of the ESDL.

The ESDL is novel and unique in its approach to focus on the fusion of global multivariate data streams and thus enables an simultaneous exploration of many facets of the Earth system. Therein, the ESDL is well equipped to face the challenges in the upcoming era of machine learning. Therefore, the ESDL and this descriptive article is an important contribution to the Earth system sciences and possibly to a wider community.

The manuscript is well structured and written. However, I have a few minor points of criticism that need consideration, before I can recommend this manuscript for publication.

1 General Comments:

1.1 *A large part of this article deals with the mathematical formulation and technical implementation of data cubes. However, you completely miss out on the technical description of the processing of the various data streams and how you treat uncertainty in the ESDL. For example, how do you treat the provided uncertainty estimates / quality flags of the individual data products? How does this effect the remapping / resampling algorithms? Is the ESDL capable to take error propagation into account? Could you provide a flow-chart to illustrate the procedure of how you incorporate data streams?*

1.2 *The second case study on the intrinsic dimension(s) of land surface variables clearly demonstrates the usefulness of the ESDL and thereby supports the title of the manuscript. Also the first case study corroborates the statement that multivariate dynamics can be better studied with data cubes, less convincingly though. However, I am not convinced that the third study really supports the need for the multivariate approach in ESDL. Here, you basically analyze only two variables (ecosystem respiration and temperature), which one could easily do with any other tool not based on data cubes. Can you make more clear why this case study supports the claim in the title?*

1.3 *The ESDL really lives on the various data streams. Many researchers have their specific datasets which they would want to analyze alongside the data streams provided in the ESDL environment. How do you enable the usage of external datasets? What are the disk usage constraints for each individual user? Is it possible to stream data (e.g. in Zarr format) from external data storage? If incorporating own datasets constitutes a complicated endeavor, researchers might be hesitant to use ESDL.*

2 Specific comments:

2.1 *L36: You cannot only analyze the state, but also the change of the system using ESDL, right?*

2.2 *L35: You start the paragraph claiming that we are well prepared in terms of data availability, but here you say there are access barriers. Maybe the term "availability" is not accurate here. A huge amount of data are collected, but they are not necessarily available for science. So, you could start this paragraph saying that we are well-prepared in terms of data collection.*

2.3 *L70: I suggest to remove 'we believe ...' and phrase the sentence: Due to its .... interface, the ESDL is well-suited ....*

2.4 *L86: This sentence ("However, ...") is somewhat complicated to grasp. I guess you want to motivate why "variable" should be treated as an additional dimension. Please revise this sentence.*

2.5 *L90: Here, the subscript of Y denotes the different variables k and the superscript denotes the different domains j - this is not consistent with the definition of Y in LL85-87.*

2.6 *L98: If the dimensions for only one grid point are dropped, do you not lose information? For example, a point measurement with certain lat and lon coordinates and a timestamp cannot be represented without losing the coordinate information, right?*

2.7 *L100: Here you use a math symbol / notation (large $\times$) to describe a cartesian product, which I have rarely seen before. Please mention that large $\times$ refers to cartesian product.*

2.8 *L105: Please define NA - does it refer to "not available", thus missing data?*

2.9 *L111: In the modelling community data cubes can be used to represent large ensembles, thus I suggest to also list "ensemble member" as another relevant dimension.*

2.10 *L214: If I may suggest to contact the maintainers of the Integrated Climate Data Center (ICDC, https://icdc.cen.uni-hamburg.de/daten.html) at the University of Hamburg. They do a very good job in collecting, processing (e.g. remapping, quality assurance), and maintaining/updating data of any kind relevant for Climate / Earth system sciences - provided in netCDF, which can easily be converted to Zarr. I assume one could join forces and build an even more comprehensive ESDL.*

2.11 *L219: "...given of their..." does not sound correct. I suggest to omit 'of' and write "... been ingested given their recurrent...".*

2.12 *LL237-247: This part explains some functions of the ESDL.jl toolbox. At this point, this information is not necessarily important for the reader. Maybe it is enough to refer to ESDL.jl documentation and the case studies which are accompanied with code - as you do it for the python implementation of ESDL.*

2.13 *L290: 'an univariate time series'*

2.14 *L291: I suggest the following corrections and revisions: "If one stored the same data cube with complete time series contained in one chunk, read operations could perform much faster."*

2.15 *L303: Delete first occurrence of 'behaviour'. I would also omit 'time', since it is redundant in the combination with 'long-term', i.e. 'long-term system behaviour in time'*

2.16 *L317: Please capitalize "northern" or use lowercase consistently for all occurrences of "northern" and "southern hemisphere".*

2.17 *L320: Please capitalize "fig" or use lowercase consistently for all occurrences.*

2.18 *L321: "Southern Hemisphere" in singular.*

2.19 *L348: I suggest using present tense when describing what you did in your study and using past tense when describing what others did before you, thus 'In our application, we follow this approach...'. Please use tenses consistently across the paper.*

2.20 *L352: Where do you introduce all the acronyms? Please provide the written-out terms here or refer to the table in the Appendix.*

2.21 *L353: I suggest to delete "the latter two", since it is not needed.*

2.22 *L361: I suggest to use the term "seasonal cycle" in singular.*

2.23 *L374: You use 'essentially', so delete 'only' in '... driven essentially by solar forcing only...".*

2.24 *L377: I recommend to use 'complex' instead of 'complicated'.*

2.25 *LL377-379: This sentence ('Zooming...') is somehow complicated to read and grasp. Maybe you can split the sentence and provide more information why it is important/interesting to focus on the northern regions of South America.*

2.26 *L382: Please use 'land surface' or 'land-surface' consistently throughout the paper.*

2.27 *L392: Here you put R_{eco} in italic letters and earlier (e.g. L352) you don't. Please use math notation consistently and follow the conventions explained in the Copernicus LaTeX template.*

2.28 *L402: Eq. (15) is incomplete. The minimal output dimensions are 'para' and 'time'. Where is the time term in the equation? Also, correct typo 'par' to 'para'.*

2.29 *L405: Please check for consistent usage of 'high-latitude, high-dimensional, high-resolution, etc.', so, with or without hyphen.*

2.30 *L418: Correct typo 'supporting materials'.*

2.31 *L426: Please check for consistent usage of 'semi-arid' versus 'semiarid' (L428).*

2.32 *L461: Where are these simplifications described in detail?*

2.33 *L464: Can 'would be' replaced by 'is'?*

2.34 *LL465-469: Please consider my comment again w.r.t. to contacting the ICDC maintainers (Comment 2.10)!*

2.35 *L510: Delete one 'several'.*

2.36 *L512: Maybe better '..., with no claim to completeness, ..." or '..., without claiming completeness, ...'.*

2.37 *L525: Please explain shortly 'kaggle' or provide an URL / reference.*

2.38 *L530: Delete one 'the'.*

2.39 *L535: The straightforward implementation of ESDL to handle / analyze CMIP multi-model but also the emerging grand ensembles of several hundreds of simulations is a key strength in the ESDL approach, in my opinion. I suggest to promote this aspect more strongly and include respective keywords, such as 'multi-model ensemble', 'large' or 'grand ensemble' in the abstract.*

2.40 *L557: Delete 'on'.*

2.41 *L557: "Dimension" in plural.*

2.42 *L594: The last sentence of the conclusion section is somewhat cumbersome. Can you boil down or split this sentence? As a reader, I expect the last sentence of the paper to be a strong and precise statement.*

2.43 *Figure 2: This is certainly an appealing visualization, however, it does not convey much information. Maybe this is also the reason why do not reference this figure anywhere in the text. I suggest to reconsider if this figure is really needed or if the url to the animation or providing the animation as ESD asset accompanying the article is sufficient. If the figure is needed, than I suggest to make some modifications as illustrated in Figure 1 and add a legend so that the figure is understandable without studying the details in the caption.*

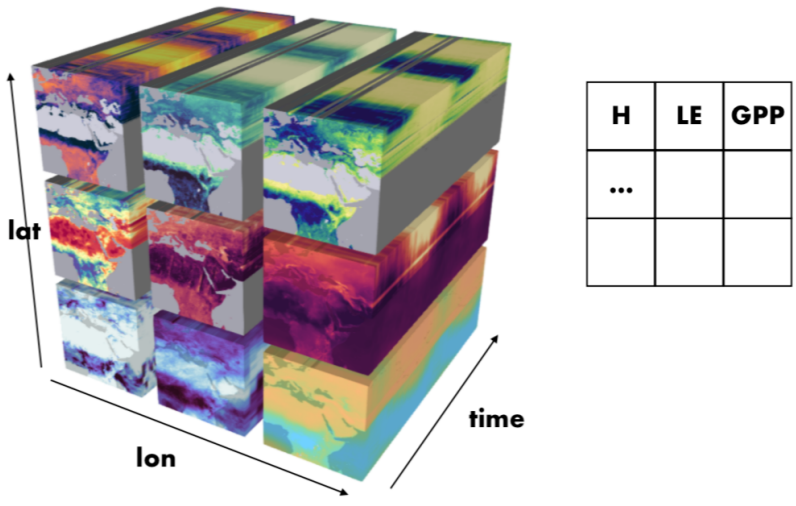

Figure 1: Modifications for Figure 2

2.44 *Figure 3: The actual values represented by the gray color-scale are not visible at all and are thus redundant in the current visualization. I suggest to include labels for actual values in the polar plot, as described here: https://matplotlib.org/gallery/pie_and_polar_charts/polar_legend.html#sphx-glr-gallery-pie-and-po*

2.45 *Figure 4: Please explain why there are no data for the Arctic.*

2.46 *Figure 5: Better vertically center the y-label.*

2.47 *Figure 6: Please correct typos: "the latter reduces .... and leads ..."*